Reactivity in panel studies and its consequences for testing causal hypotheses

van der Zouwen, J.; van Tilburg, T.G.

published in
Sociological Methods and Research
2001

DOI (link to publisher)
10.1177/0049124101030001003

document version
Publisher's PDF, also known as Version of record

Reactivity in Panel Studies and its Consequences for Testing Causal Hypotheses

JOHANNES VAN DER ZOUWEN and THEO VAN TILBURG

Sociological Methods & Research 2001 30: 35
DOI: 10.1177/0049124101030001003

The procedure of standardized repeated measurement, as used in panel studies, may hamper the quality of the data, due to the potential "reactivity" of survey interviewing on respondents' attitudes and behavior. In case respondents are interviewed in subsequent waves by different interviewers, differential interviewer effects may occur. These threats to data quality are illustrated with data from a longitudinal study among 2,819 older adults, conducted in the Netherlands. From an analysis of 100 interview protocols, it appears that the behavior of the interviewers has a significant impact on the data obtained. On one hand, interviewers seem to adjust their interviewing strategy to a norm regarding a "normal" personal network and, on the other hand, to a norm about the appropriate interviewing time. Suggestions are formulated to prevent misestimating actual change within respondents over time, leading to incorrect conclusions about causal relationships.

Reactivity in Panel Studies and Its Consequences for Testing Causal Hypotheses

JOHANNES VAN DER ZOUWEN
THEO VAN TILBURG
Vrije Universiteit, Amsterdam

Causal hypotheses can be tested in different ways: in true or quasi-experimental designs, but also in panel studies. Panel studies require successive measurements of the same participants with the same instruments, for example, repeated interviewing of the same respondents by (preferably) the same interviewers, asking the same questions. However, this procedure of standardized repeated measurement may hamper the quality of the data, due to the potential "reactivity" of survey interviewing on respondents' attitudes and behavior, especially the occurrence of Socratic effects, fatigue effects, and the memory effects regarding responses given in subsequent interviews.

AUTHORS' NOTE: *This article is based on data collected in the context of the Living Arrangements and Social Networks of Older Adults (LSN) and Longitudinal Aging Study Amsterdam (LASA) research programs. These programs are conducted at the Vrije Universiteit in Amsterdam and the Netherlands Interdisciplinary Demographic Institute in The Hague and are supported by the Netherlands Program for Research on Aging (NESTOR) and the Ministry of Health, Welfare, and Sports.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 30 No. 1, August 2001 35-56

When respondents are interviewed in subsequent panel waves by different interviewers, which often cannot be avoided in practice, differential interviewer effects may occur. Even if these effects on the measurements at different points in time are only random, they will lead to less accurate estimates of actual change. The degree of unreliability of each of the separate measurements will be amplified if differences between the measurements at succeeding points in time (difference scores) are calculated.

We will concentrate on the occurrence of reactivity and differential interviewer bias as threats to the validity of conclusions about change as inferred from repeated measurements in standardized panel studies. We will focus our analysis on a panel study designed to investigate the personal networks of older adults. The question we attempt to answer is, To what degree, under which conditions, and how do reactivity and differential interviewer bias occur in this panel study, and is it possible to prevent these threats to the validity of causal inferences?

REACTIVITY IN PANEL STUDIES

FORMS OF REACTIVITY

In the literature, reactivity in panel studies (i.e., the effects of the observation at time T_1 on the data obtained concerning the same participants and the same variables at subsequent time points T_2, T_3, etc.) is indicated by different terms: repeated measurement effect, interview effect, panel effect, panel conditioning, and time-in-sample bias. In the remainder of this article, we will use the term *panel effect* (Cantor 1989). To better understand how this panel effect operates, one has to distinguish between the following possibilities or forms of panel effects. The observation at T_1 may have an effect at T_2 on (1) the value of variable Y, (2) the score on one or more of the indicators of Y, (3) the relationship between Y and its indicators y_i, or (4) the measurement errors of the indicators of Y.

PANEL EFFECTS ON VARIABLE Y

In many cases, variable *Y* is an attitude. Repeated measurement may affect the attitude of the respondent by "raising consciousness" (Waterton and Lievesley 1989) due to the previous measurement, especially if the attitude is not very salient to the respondent at the time of the first interview (Bridge et al. 1977). Veroff, Hatchett, and Douvan (1992) observed a negative effect of repeated measurement on marital satisfaction of couples in the first years of their marriage, an effect that changed into a positive one for later years of their marriage. Glock (1952) suggested that weak attitudes about political issues may crystallize ("freezing") due to repeated interviewing. However, Waterton and Lievesley (1989) did not find much empirical evidence for the occurrence of freezing. Clausen (1968) observed that the probability to participate in an election increased after being interviewed about political issues. He assumed that the interview enhanced the interest of the respondent in politics. This stimulus hypothesis was adopted in research conducted by Yalch (1976) and Traugott and Katosh (1979). In a study by Brannen (1993), another instance of an effect of repeated measurement was observed: Women who were repeatedly interviewed during maternity leave differed with respect to their willingness to return to the labor market from those who were interviewed only once.

We just identified some apparent conditioning effects on attitudes and behavior. However, in other studies about conditioning effects of repeated measurement, these effects turned out to be unsystematic, small, and hardly distinguishable from changes in response behavior (Silberstein and Jacobs 1989). The conditioning effects appeared to be less threatening to data quality than recall error (Holt 1989). In studies on health condition and medical consumption (Corder and Horvitz 1989) and on consumer behavior (Sobol 1959), conditioning effects were even absent. Klein and Rubovits (1987) observed that respondents interviewed in various waves of a panel study reported as many stressful events as respondents who were interviewed only once.

This short overview of the literature shows that conditioning effects sometimes appear and sometimes do not, without a clear indication of the conditions under which these effects occur.

PANEL EFFECTS ON INDICATORS OF VARIABLE Y

A respondent, interviewed for the second time, might remember the answers he or she gave in the first interview and repeat this answer, irrespective of whether the answer is still correct or not (O'Muircheartaigh 1989). In this way, memory effects could enhance reactivity. However, the occurrence of such a memory effect of the first interview on a second interview, conducted several months later, is very unlikely, because Scherpenzeel (1995) observed that memory effects within the same interview were already extinguished within a time span of 20 minutes.

PANEL EFFECTS ON THE RELATIONSHIP BETWEEN VARIABLE Y *AND ITS INDICATORS*

If the relationship between variable Y and its indicators y_i changes, then the score on y_i may change, even if the value on the variable Y has remained the same. This means that there is a change in the degree to which the reactions of the respondents to these items of the questionnaire are informative about variable Y, implying a change in data quality. In the literature, we observed contradictory statements about the effect of repeated measurement on data quality.

On one hand, a worsening of data quality is observed or at least supposed: Respondents get bored and fatigued and will become less motivated to search for the correct answer. Moreover, respondents find out that certain answers (e.g., positive answers to the question "Have you ever . . . ?") lead to more follow-up questions than others. In the long run, respondents will then try to avoid these follow-up questions as much as possible, which will lead to negative answers and consequently to underreporting. This worsening of data quality will especially occur within long interviews (Bailar 1975, 1989; Cantor 1989; Corder and Horvitz 1989).

However, other researchers observed an improvement of data quality in repeated interviews. This might be explained by the fact that respondents have gradually received a more precise image of the meaning of the questions and consequently answer them more adequately (Bailar 1989; Cantor 1989). For example, Traugott and Katosh (1979) obtained higher proportions of accurate answers to

questions about participation in elections in repeated interviews. Moreover, respondents may become more convinced about the importance their answers have for the researcher, even if their opinion is not as yet quite firm. Waterton and Lievesley (1989) indeed observed a decrease of the proportion of "don't know" responses to opinion questions in repeated interviews. For retrospective questions, Bailar (1989) observed that recall errors due to telescoping decreased considerably after the second and following interviews.

Porst and Zeifang (1987) supposed that one reason for the improved data quality of repeated interviewing is that respondents know that next-time questions about the same topics will be posed again and that, therefore, they do the best they can to give a correct answer. Waterton and Lievesley (1989) assumed that being interviewed repeatedly, especially by the same interviewer, will make the interview itself less threatening to the respondent, thereby decreasing the need to give socially desirable and artificially consistent answers. The latter phenomenon would lead to a decrease of interitem correlations. This is exactly the reverse from what one would expect if a so-called Socratic effect (Jagodzinski, Kühnel, and Schmidt 1987) occurred, where the repeated interviewing is assumed to increase the coherence of the attitude and consequently to increase the interitem correlations (see also Feldman and Lynch 1988).

PANEL EFFECTS ON THE MEASUREMENT ERRORS OF THE INDICATORS

Part of the measurement error is respondent specific. Presser and Traugott (1992) showed that specific respondents are inclined to give socially desirable responses. Their answers to the same questions y_i posed at succeeding points in time will remain the same, even if their position on variable Y changes. In this case, the measurement error is not changing due to repeated measurement, but the magnitude of the measurement error may change during the repeated measurement.

Anyway, we like to repeat the remark of Waterton and Lievesley (1989) that we will never be able to infer from changes in responses to items y_i at subsequent time points, whether the position of the respondent on variable Y has changed ("real" change) or that the change d_y is due to change in response behavior or, rather, measurement error.

PANEL EFFECTS AND CHANGES IN REPORTED PERSONAL NETWORK SIZE

REACTIVITY AND DIFFERENTIAL INTERVIEWER EFFECTS IN PRACTICE

The overview of the few empirical studies on panel effects does not lead to firm conclusions about the effect of repeated measurement in panel studies on the measurement of the latent variable itself or on the validity and reliability of its indicators (*reactivity* for short). There is some empirical evidence for the occurrence of conditioning effects in studies of political participation, the occurrence of memory effects seems unlikely, and there are contradictory findings regarding the effects of repeated interviews on the quality of the responses obtained. The literature on differential interviewer effects is even scarcer, let alone that it provides explanations for these effects. For that reason, we analyzed panel data to study the mechanisms of reactivity and differential interviewer effects. Data were available from wave 1 and wave 2 of a longitudinal survey on the personal networks of older adults. We selected this study because we knew that the researchers had chosen for a panel design to assess as well as possible the causal relations between network size; its determinants, such as age and health; and its effects, such as feelings of loneliness. Another reason for selecting this study was that we had access not only to the usual panel data, such as the responses of the respondents to the questionnaire, but also to detailed information about what went on during the course of the interview, since we could make use of audiotapes of the interviews and verbatim transcripts of these tapes. This made it possible to investigate (at least partly) to what extent the data collected in this panel study were affected by the process of data collection itself.

DESCRIPTION OF THE DATA

Respondents

Personal interviews were conducted in 1992 (T_1) with 3,805 respondents who participated in the Living Arrangements and Social Networks of Older Adults (LSN) research program (Knipscheer et al.

1995). This program used a stratified random sample of men and women born between 1908 and 1937. The oldest individuals, and in particular the oldest men, were overrepresented in the sample. The sample was taken from the population registers of 11 municipalities in the Netherlands. Respondents were interviewed in their homes, and personal computer assistance (CAPI) was used in the data collection. The interviews mainly covered demographics, the personal network, loneliness, and life event history. In 1992-1993 (T_2), a follow-up was carried out in the context of the Longitudinal Aging Study Amsterdam (LASA) (Deeg and Westendorp-de Serière 1994). The LASA interviews covered a wide range of topics relating to physical and cognitive health, and social and psychological functioning. Of the T_1 respondents, 3,107 (82 percent) participated in the follow-up. Data of subsequent waves, conducted in 1995-1996 (T_3) and in 1998-1999 (T_4), are not included in the analyses. The interval between T_1 and T_2 averaged .86 years. The interviews lasted between one and one-half and two hours. The networks of 2,819 respondents were delineated in both waves using the same method. Van Tilburg (1998b) reported the details of the design and of the observations.

Interviewers

The interviewers were trained for four days. They were told the general interviewing rules and practiced putting them into effect with role-playing. Videotapes with common interviewing situations and mistakes were shown and discussed. All the sections of the questionnaire were discussed, and particular attention was devoted to the large number of different routings and to deviating question formats. Sections of the questionnaire were practiced with another interviewer acting as the respondent. A complete pilot interview with someone (age 55 to 89) known by the interviewer was conducted in the respondent's home. The experiences with these interviews were discussed. All procedures concerning contact with the respondents and with the research team were written down or were part of the computer program and were clarified during the training. During the data collection, the supervisors listened to tape recordings of the interviews. They weekly discussed interview style, suggestive questioning, handling difficult situations, administrative matters, and so forth with the

interviewers. Furthermore, other training sessions were held, including, among others, a meeting to discuss all kinds of interviewing problems.

At T_1, there were 87 interviewers who interviewed on average 52 respondents. For T_2 and T_3, there were 43 interviewers who interviewed on average 72 respondents. In most of the cases, the respondents were interviewed at T_1 and T_2 by different interviewers (n = 2,450).

THE MEASUREMENT OF PERSONAL NETWORK SIZE

The main objective of this study was to identify a network that reflected the socially active relationships of the older adults in the core, as well as the periphery, of the network. Several criteria were applied to the selection of a method for identifying the personal network, with regard to whom was to be included in the network. First, the network composition had to be as diverse as possible, implying that all types of relationships deserved the same chance to be included in the network. This criterion led to a domain-specific approach in the network identification, using seven formal types of relationships: household members (including the spouse, if there was one), children (including stepchildren) and their partners, other relatives, neighbors, colleagues (including voluntary work or school), fellow members of organizations (e.g., athletic clubs, church, political parties), and others (e.g., friends and acquaintances). A second objective was to include all the network members with whom the respondent had regular contact, thus identifying the socially active relationships. To avoid picking out individuals who were contacted frequently by definition (such as all the members of a club), the importance of the relationship was added as a criterion.

This domain-contact approach combines the various roles an individual plays in society with the contact frequency and the importance of the relationships as criteria for the identification of network members. For each of the seven domains, the following question was posed: "Please name the people (e.g., in your neighborhood) you have frequent contact with and who are also important to you." The interpretation of the criteria was left to the respondent. Only people above

age 18 could be nominated. The maximum number of names was set at 80, but no one reached this limit. The design of the measurements was the same for each of the four observations, giving network members identified in a previous observation and others the same chance to be identified in later observations. The total network size was computed as the number of individuals identified. Partial networks were identified by the type of network member, that is, child (including stepchildren), child-in-law, sibling, sibling-in-law, other relative (e.g., parent, grandchild, aunt, uncle, nephew, niece), friend, neighbor, or other nonrelative (e.g., acquaintance, colleague, or fellow member of an organization) of the respondent. The sum of the partial network sizes equals the total network size. However, the spouse or partner was not included in the computation of the partial network sizes.

For detecting changes in the network composition, the names of all the network members identified in different observations were compared and, if possible, linked. To enhance this, during the network delineation at T_2, the interviewer had available a list of the network members identified at T_1 and was requested to link the identified network members with those identified at T_1. The interviewers were instructed to use the T_1 information for matching purposes only, not to discuss for what reason a network member identified at T_1 was not identified at T_2. Based on the data concerning the links, the joint network size at two observations could be broken down into three parts: the number of network members identified at both observations ("stable"), the number of members not identified at the first observation ("gained"), and the number of members identified at the first observation but not identified at the second observation ("lost"). The T_1 and T_2 network sizes and the numbers of stable, of lost, and of gained network members are, of course, not independently measured. Therefore, in the analyses, three variables to be explained will be used: the network size at T_2, the proportion of T_1 members not identified at T_2 ("lost members"), and the number of T_2 members not identified at T_1 ("gained members"), as proportion of the T_1 network size. The T_2 network size correlated –.20 with the proportion of lost members and .14 with the proportion of gained network members; the latter two correlated –.37.

As reported by van Tilburg (1998a), there were large interviewer effects in the measurement of the personal network size at T_1: the

intraclass correlation ρ_{int} was .21 and decreased only marginally when it was controlled for a large number of respondent characteristics. To determine whether these interviewer effects were specific for this particular set of questions, the interviewer effects on the network delineation were compared with those on the scores on a standardized 11-item scale on loneliness. For the loneliness scores, computed as the sum of the item scores, ρ_{int} was .02. This indicated that the interviewer effects were large for the measurement of network size and only small for the measurement of loneliness.

This is a remarkable outcome because interviewer effects are usually stronger on "attitudinal information" than on "behavioral information" (Sudman and Bradburn 1974). We consider the following explanations. First, the questions about the personal network are indeed factual questions, but they are more difficult to answer than most other factual questions. The network delineation questions belong to the category of open questions, requiring a "stock taking" of the social environment and subsequently categorizing the set of personal relationships into various domains (which persons belong to what domains?), as well as recalling social interactions that have taken place during the last couple of years. Second, the answers to questions about personal relationships have a normative loading: You ought to have good relationships with your children and other family members, and moreover, the number of friends you have symbolizes in a way your friendliness and sociability. Third, the seven social domains are not mutually exclusive. A colleague or a neighbor may also be a good friend, so it may not be clear in which domain this person should be identified. Furthermore, the criteria "frequent contact" and "important for you," mentioned in the questions, are sometimes difficult to apply, not only for the respondent but also for the interviewer, the more so while these criteria have a different meaning for different domains and different forms of personal contact (e.g., personal visits versus telephone calls). Each of these question characteristics may lead to specific interviewer behaviors, which may contribute to the large interviewer effect (van Tilburg 1998a).

Van der Zouwen and Dijkstra (1988) observed that during an interviewing campaign, interviewers gradually change their interviewing strategies: They "learn" from experience gained in preceding

TABLE 1: Means and Standard Deviations of Total and Partial Network Sizes (N = 2,819) as Measured at T_1 and T_2

	T_1		T_2	
	M	SD	M	SD
Total network size	14.6	9.8	13.9	8.3
Partial network size				
Children	2.4	1.8	2.6	1.9
Children-in-law	1.5	1.6	1.6	1.7
Siblings	1.2	1.7	1.4	1.6
Siblings-in-law	1.5	2.5	1.4	2.3
Other relatives	1.4	2.5	1.2	2.0
Friends	1.5	2.8	1.5	2.5
Neighbors	1.8	2.3	1.6	1.9
Other nonrelatives	2.5	3.9	1.9	2.9

NOTE: T_1 = Time 1; T_2 = Time 2.

interviews how to deal with the often contradictory goals of establishing a good "rapport" with the respondent and of obtaining high data quality within severe time limits. With regard to the network delineation reported here, there is an indication for such a learning effect. For the T_1 interviews, a positive and significant association was observed between reported network size and the sequence number of the interview held by a particular interviewer (van Tilburg 1998a).

We expect that the estimated degree of change in network size will be influenced by wording effects due to the partial indefiniteness of the meaning of the criteria "frequent" and "important" mentioned in the questions, the normative character of the conceivable responses, and the (changing) differential effects of the interviewers on the responses. This expectation can be partially checked by comparing the data concerning network size of respondents collected at subsequent waves. We will concentrate the analysis on a comparison of the T_1 and T_2 observations.

CHANGES IN REPORTED NETWORK SIZE

To test the causal hypotheses regarding causes and effects of changes in network size, it is of vital importance that changes in network size are accurately estimated. A rough indication of these

changes is presented in Table 1. The average total network size decreased between T_1 and T_2 ($t_{(2,818)} = 4.6$, $p < .001$, $r = .64$). The same applies to the standard deviation for nearly all the partial networks.

It appears that the smallest networks (0 to 7 members) observed at T_1 became larger (from an average of 4.7 to 7.7) and the largest (19 or more members) became smaller at T_2 (from an average of 27.1 to 20.8). van Tilburg (1998b) showed that this phenomenon occurs especially with the domains "other relatives" and "nonrelatives," precisely those domains that appeared to be the most sensitive to interviewer effects. Part of the observed changes in the total network size between T_1 and T_2 thus might be due to the fact that most respondents (87 percent) are interviewed at T_2 by an interviewer different from the one at T_1. Among respondents interviewed at T_1 and T_2 by the same interviewer, the network size was stable (13.8 and 14.1, respectively; $t_{(368)} = -.8$, $p > .05$, $r = .70$), whereas the network size decreased among the other respondents (14.7 and 13.8, respectively; $t_{(2,449)} = 5.2$, $p < .001$, $r = .63$). The question of how much the differences between network size at T_1 and T_2 for those respondents who are interviewed at T_2 by another interviewer are caused by reactivity of the (repeated) measurement, or by differential interviewer behavior, cannot be answered with the data in the data matrices. Thus, we have to analyze the interviews themselves, that is, we have to perform a protocol analysis of the audiotaped interviews.

ANALYSIS OF INTERVIEW PROTOCOLS

THE DESIGN OF THE PROTOCOL ANALYSIS

The part of the interviews held at T_2 that was related to the questions about network size was transcribed into verbatim protocols. Not all interviews were transcribed but only those that satisfied the following five criteria: (1) Of the respondents involved, there are also interviews available of the preceding (T_1) and subsequent (T_3) wave; (2) these three interviews were conducted by three different interviewers; (3) during this study, the partner status of the respondent (has he or she a partner or not) did not change; (4) the tapes were (technically) in good

shape; and (5) the respondents were interviewed by interviewers who had held at least 15 interviews.

The resulting protocols, about 600 together, amount to 1,400 pages of text. For the present analysis, we took from these protocols a random sample of $N = 100$. The selected protocols were coded according to three aspects:

1. *Any reference made by the respondent to the foregoing interview.* Our expectation is that this will lead to a more stable network, that is, less members lost and gained between T_1 and T_2.
2. *The way in which the interviewer refers to the preceding interview with this respondent.* We expect that if the interviewer explicitly uses information about the respondent's network collected during the preceding interview, this will also lead to a more stable network.
3. *Efforts of the interviewer to steer the reporting of the respondent, especially by incorrectly applying the delineation criteria, either by restricting or broadening them.* Our expectation is that this interviewer behavior will lead to more losses, respectively more gains, at T_2.

SOCRATIC EFFECTS AND MEMORY EFFECTS

Repeated measurement might have a so-called Socratic effect (Jagodzinski et al. 1987) on variable Y or a memory effect on the indicators of Y, that is, y_i. If these effects had occurred in the present study, one would expect that respondents interviewed at T_2 would have referred to their answers to the same questions given in the foregoing interview or at least to the fact that they already had answered these questions. The analysis of the 100 interview protocols showed that in 84 cases, the respondent did not make any reference to the preceding interview. Five respondents referred to the preceding interview only after the interviewer had mentioned that interview. In only 11 cases, respondents referred on their own initiative to the earlier interview, for example, by mentioning that they already had answered these questions before. Whether a respondent made a reference to the previous interview, controlled for the T_1 network size and the length of the interval between T_1 and T_2, was not significantly associated with the T_2 network size nor with the proportion of lost and gained network members ($F_{(2,95)} = 1.3, F_{(2,94)} = 1.6, F_{(2,94)} = 1.4$, respectively). This means that our first expectation is not confirmed by the data.

Of course, one never knows for sure whether the preceding interview had an effect on the reporting of network size in the subsequent one. But the fact that so few respondents referred to the earlier interview makes it unlikely that in this study, reactivity, in the form of Socratic effects or memory effects, played an important role. This outcome is in line with the results of Corder and Horvitz (1989).

REFERENCES TO THE PREVIOUS INTERVIEW MADE BY THE INTERVIEWER

In 40 of the 100 cases, the interviewers explicitly made reference to the previous interview, by telling the respondents that they had information projected on the computer screen about their answers given in the previous interview. In 33 of these cases, the interviewers compared the present responses with the ones given earlier. Sometimes this comparison was only intended to find out whether the person mentioned at T_1 is the same as a person now mentioned with a somewhat different name ("Is that the same person as Mary?"). However, in many cases, the interviewers tried to compare for each domain the members mentioned now and those mentioned in the previous interview, which in some cases even led to an effort to adjust the present reporting to the previous one ("Last time you also mentioned" or "You did not mention . . . last time").

Such an explicit reference, made by the interviewer, to the answers given in the previous interview indeed has a strong effect on the present responses. The results of a regression analysis showed that, controlled for the T_1 network size ($\beta = .39$, $p < .001$) and the interval between the T_1 and T_2 observations ($\beta = .02$, $p > .05$), the proportion of T_1 members not identified at T_2 (lost members) was negatively associated ($\beta = -.25$, $p < .01$) with whether the interviewer explicitly referenced to T_1. In explaining the T_2 network size and the proportion of gained network members, no significant effects were observed.

It can be concluded that in a considerable number of cases, the previous measurement has an effect on the subsequent one, not through some cognitive processes within the respondent but through the behavior of part of the interviewers, behavior that was surely not intended by the researchers. That some respondents seem to have lost

more members in their network than others is thus partly an artifact, due to differential interviewer bias.

THE APPLICATION OF THE NETWORK CRITERIA

The criteria of "frequent contact" and "important contact," as applied in the network delineation, may have resulted in different interpretations. It appears from the transcripts of the interviews that this freedom of interpretation led to many discussions between respondents and interviewers about the proper application of these criteria. During these discussions, the interviewer has the opportunity to influence the application of these criteria by the respondent. This may lead to more appropriate reporting but also to such a highlighting of the criteria that the network size reported at T_2 becomes smaller than the one reported at T_1.

According to the instructions, in 60 cases, interviewers only repeat both criteria. However, in 25 cases, we observed clear attempts by the interviewer to curb respondents' reporting. One may distinguish two subcategories here: (1) sharpening the criteria, for example, changing "frequent contact" into "intensive contact" or changing "important contact" into "people with whom you share life's joys and sorrows" (12 cases); (2) ranking the potential subject according to contact frequency and importance, and selecting from this list only those with whom the respondent has the most frequent contacts and/or are the most important for him or her (13 cases).

In interviews in which the interviewer uses a selection strategy, the number of network members gained at T_2 is larger than if the interviewer does not use such a strategy. The results of a regression analysis showed that, controlling for the T_1 network size ($\beta = -.42, p < .001$) and the interval between the T_1 and T_2 observations ($\beta = .15, p = .11$), the proportion of gained members was positively associated ($\beta = .16$, $p = .11$) with the application by the interviewer of a selection strategy (in explaining differences in the T_2 network size and the proportion of lost members, much smaller effects were observed). At first sight, this positive effect (although it is small and not significant) is not only contrary to our expectations but also unexplainable. Therefore, we have to continue the analysis.

We saw that between T_1 and T_2, the smallest networks became larger, and the largest ones became smaller. Apart from the explanations for this phenomenon offered by van Tilburg (1998b), we can also suppose that the decision of the interviewer at T_2 to influence the respondent's reporting of his or her network is based on the information about the network size reported at T_1—information available to the interviewer at T_2—and on the experience gained by the interviewer in earlier held interviews. The interviewers may have learned from the previous interviews about what network size they can expect in the present interview, and they may have also learned that large networks require a lot of time and effort to be reported in detail. We expect that interviewers use their recent experience, that is, the maximum network size observed in one of the three T_2 interviews just preceding the present one, as well as information about the size of the network of the respondent reported at T_1, to adjust their interviewing strategy.

It appears that if in one of the three previous interviews at T_2 a large network is identified, this experience affects the present interview. In the explanation of the proportion of lost members at T_2, a negative effect of this maximum was observed ($\beta = -.17, p < .10$; the beta of the T_1 network size changed to .40 and the beta of whether the interviewer referenced to T_1 was not affected). Interviewers seem to probe further for additional network members as long as the reported network size is below the norm of a proper network. The maximum network size reported in one of the three previously held interviews also has a positive effect on the number of members gained at T_2 ($\beta = .11, p = .24$; the beta of the T_1 network size changed to –.43 and the beta of whether the interviewer applied the selection strategy changed to .17, $p < .10$).

To check this line of thought, we analyzed whether the T_1 network size and the maximum network size in the previous three interviews could explain the occurrence of the selection strategy. The results of a logistic regression analysis showed that there was indeed evidence supporting the hypothesis: Interviewers were less likely to apply the selection strategy when in the three previous T_2 interviews large networks were identified ($R = -.12, p < .10$) and were more likely to apply the selection strategy when the T_1 network size was large ($R = .21, p < .05$) and when the respondent was identifying a large number of network members at T_2 ($R = .08, p = .12$).

The processes just mentioned indicate an interplay between the variables regarding available information about the network size of this particular respondent, the experiences gained by the interviewer in previously held interviews, and the choice of the interviewer for a certain interviewing strategy. On one hand, interviewers seem to adjust their interviewing behavior to a norm regarding a "normal" network and, on the other hand, to a norm about the appropriate interviewing time.

CONCLUSIONS

One of the main criticisms regarding the survey is that causal hypotheses can only indirectly be tested by survey data. For a more powerful test of causal hypotheses, one needs multimoment observations like those collected with, for example, panel studies. However, panel studies have their own methodological drawbacks, as appeared in the short summary of the literature on panel conditioning. The preceding observation may affect the subsequent observation in various ways: The latent variable itself may be affected, or its indicators, or the relationship between the latent variable and its indicators, or the reliability of the indicators.

In the panel study discussed in this article, there was hardly an indication of the occurrence of panel conditioning within the respondents. What really caused worry was the fact that the behavior of the interviewer had such a large impact on the dependent variable, network size. As in most panel studies, it turned out impossible to have all respondents be interviewed during all the waves of the data collection by the same interviewers. So we could expect to find differential interviewer bias.

But that was only part of the story. From a detailed analysis of interview protocols, we learned that interviewers apply some kind of steering behavior: They can enhance reported network size by further probing and curb the network size by using a selection strategy in which criteria are made more strict than meant by the researcher or by asking respondents to only mention those people who come closest to these criteria. This steering behavior is based, on one hand, on

information the interviewer possesses about the network size reported in the preceding interview with this particular respondent and, on the other hand, on the experience the interviewer gained with reported network size in the three previously held interviews. This means that the second observation of the network size was affected both by the previous observation of this variable and by the expectations or norms interviewers develop during their present data collection. This feedback loop between expected network size and interviewer behavior leads to an underestimation of the variance in the dependent variable network size at a particular moment and to some sort of "regression to the mean" over time. To summarize, we observed reactivity of the measurement, or panel conditioning, but this did not take place via cognitive processes within the respondent's mind but via the task-related behavior of the interviewer.

It is appropriate to mention here that the interviewers were instructed and supervised at least as strictly as in other panel studies. So, we do not think that the outcome of our study is only a regrettable exception. If one needs interviewers for data collection because the information to be reported is complex and if this information has a normative "loading," then we can expect to meet these methodological problems in any comparable panel study.

One can think of various actions to at least decrease the negative effects of differential interviewer bias and panel conditioning. The first one is to reduce the workload of the interviewers. The second one is a very careful testing of the more complex parts of the questionnaire using "cognitive" interviews and interaction analysis (Schaeffer and Maynard 1996; van der Zouwen and Dijkstra 1996). Another action to be taken is even stricter instruction of the interviewers and more frequent supervision of the interviewers, to see whether they still stick to the rules for further probing and for explaining the questions of the questionnaire. These actions are important for all surveys, cross-sectional and panel studies alike. However, they are the more urgent for panel studies since the observed change might be strongly biased in case of differential interviewer effects, which will affect the opportunity to draw conclusions on causality.

Furthermore, within many panel studies, interviewers have available data from previous observations because questions are asked on (the consequences of) differences between the current and the

previous situation. From the present study, it is clear that interviewers used the information from the previous observation. Moreover, the usage of that information affected the data collected. The availability of data from the previous observation might enhance the data quality, for example, by its resulting in less underreporting of the network size. However, at the same time, this procedure may harm the validity of observed differences between observations since such an action to prevent underreporting cannot be taken at the first observation and the results of such an action will differ across interviewers. The supervisors of the panel study presented here recognized these effects. Consequently, the procedure changed at T_3. The questions on the matching of network members identified at the current and at previous observations were separated from the questions on the identification, and the list of network members identified at earlier observations became available to the interviewers only after the identification of the network members.

In testing a causal hypothesis empirically by analyzing panel data, the estimation of change in a variable might be biased by repeated measurement. The results of the analyses show that the assumption of later observations being independent from previous observations was violated. The data collected at previous observations influence the interviewers' expectations on data to be collected at later observations and steer their behavior in asking questions. This panel effect will threaten the validity of the repeated measurement and, consequently, of conclusions on the causal hypothesis to be tested. We recommend not providing the interviewers with information from previous observations or providing this information after the factual situation at later observations has been measured if it is necessary to have that information available. Furthermore, it was concluded that interviewers seem to adjust their interviewing behavior to a norm about the appropriate interviewing time. Generating more responses on open questions will extend the interviewing time. In an analysis of the T_1 data, van Tilburg (1998a) observed that interviewers with interviewing experience prior to the project generated relatively small networks. It was suggested that these interviewers were less open to following instructions. We recommend not selecting interviewers with prior interviewing experience.

REFERENCES

Bailar, Barbara A. 1975. "The Effects of Rotation Group Bias in Estimates From Panel Surveys." *Journal of the American Statistical Association* 70:23-30.

———. 1989. "Information Needs, Surveys and Measurement Errors." Pp. 1-24 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh. New York: John Wiley.

Brannen, Julia. 1993. "The Effects of Research on Participants: Findings From a Study of Mothers and Employment." *The Sociological Review* 41:328-46.

Bridge, R. Gary, Leo G. Reeder, David Kanouse, Donald R. Kinder, Vivian Tong Nagy, and Charles M. Judd. 1977. "Interviewing Changes Attitudes—Sometimes." *Public Opinion Quarterly* 41:56-64.

Cantor, David. 1989. "Substantive Implications of Longitudinal Design Features: The National Crime Survey as a Case Study." Pp. 25-51 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Clausen, Aage R. 1968. "Response Validity: Vote Report." *Public Opinion Quarterly* 32:588-606.

Corder, Larry S. and Daniel G. Horvitz. 1989. "Panel Effects in the National Medical Care Utilization and Expenditure Survey." Pp. 304-18 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Deeg, Dorly J. H. and Mariëtte Westendorp-de Serière, eds. 1994. *Autonomy and Well-Being in the Aging Population: Report From the Longitudinal Aging Study Amsterdam 1992-1993*. Amsterdam: VU University Press.

Feldman, Jack M. and John G. Lynch. 1988. "Self-Generated Validity and Other Effects of Measurement on Belief, Attitude, Intention, and Behavior." *Journal of Applied Psychology* 73:421-35.

Glock, C. 1952. "Participation Bias and Reinterview Effects in Panel Studies." Ph.D. dissertation, Columbia University, New York.

Holt, D. 1989. "Panel Conditioning: Discussion." Pp. 340-47 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Jagodzinski, Wolfgang, Stephen M. Kühnel, and Peter Schmidt. 1987. "Is There a 'Socratic Effect' in Nonexperimental Panel Studies? Consistency of an Attitude Toward Guestworkers." *Sociological Methods and Research* 15:259-302.

Klein, Daniel N. and David R. Rubovits. 1987. "The Reliability of Subjects' Reports on Stressful Life Events Inventories: A Longitudinal Study." *Journal of Behavioral Medicine* 10:501-12.

Knipscheer, Kees C.P.M., Jenny de Jong Gierveld, Theo van Tilburg, and Pearl A. Dykstra, eds. 1995. *Living Arrangements and Social Networks of Older Adults*. Amsterdam: VU University Press.

O'Muircheartaigh, Colm. 1989. "Sources of Nonsampling Error: Discussion." Pp. 271-88 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Porst, Rolf and Klaus Zeifang. 1987. "A Description of the German General Social Survey Test-Retest Study and a Report on the Stabilities of the Sociodemographic Variables." *Sociological Methods and Research* 15:177-218.

Presser, Stanley and Michael Traugott. 1992. "Little White Lies and Social Science Models: Correlated Response Errors in a Panel Study of Voting." *Public Opinion Quarterly* 56:77-86.

Schaeffer, Nora Cate and Douglas W. Maynard. 1996. "From Paradigm to Prototype and Back Again: Interactive Aspects of Cognitive Processing in Standardized Survey Interviews." Pp. 65-88 in *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, edited by N. Schwarz and S. Sudman. San Francisco: Jossey-Bass.

Scherpenzeel, Annette 1995. "A Question of Quality: Evaluating Survey Questions by Multitrait-Multimethod Studies." Ph.D. dissertation, University of Amsterdam.

Silberstein, Adriana R. and Curtis A. Jacobs. 1989. "Symptoms of Repeated Interview Effects in the Consumer Expenditure Interview Survey." Pp. 289-303 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Sobol, M. G. 1959. "Panel Mortality and Panel Bias." *Journal of the American Statistical Association* 54:52-68.

Sudman, Seymour and Norman M. Bradburn. 1974. *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.

Traugott, Michael W. and John P. Katosh. 1979. "Response Validity in Surveys of Voting Behavior." *Public Opinion Quarterly* 43:359-77.

van der Zouwen, Johannes and Wil Dijkstra. 1988. "Types of Inadequate Interviewer Behavior in Survey Interviews: Their Causes and Effects." *Bulletin de Méthodologie Sociologique* 18:5-20.

———. 1996. "Testing (CATI-)Questionnaires Using Computer Assisted Interaction Coding." Presented at the International Conference on Computer-Assisted Survey Information Collection, December 11-14, San Antonio, TX.

van Tilburg, Theo. 1998a. "Interviewer Effects in the Measurement of Personal Network Size: A Nonexperimental Study." *Sociological Methods and Research* 26:300-28.

———. 1998b. "Losing and Gaining in Old Age: Changes in Personal Network Size and Social Support in a Four-Year Longitudinal Study." *Journal of Gerontology* 53B:S313-23.

Veroff, Joseph, Shirley Hatchett, and Elizabeth Douvan. 1992. "Consequences of Participating in a Longitudinal Study of Marriage." *Public Opinion Quarterly* 56:315-27.

Waterton, Jennifer and Denise Lievesley. 1989. "Evidence of Conditioning Effects in the British Social Attitudes Panel." Pp. 319-39 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Yalch, Richard F. 1976. "Pre-Election Interview Effects on Voter Turnout." *Public Opinion Quarterly* 40:331-36.

Johannes van der Zouwen is head of the Department of Social Research Methodology, Vrije Universiteit, the Netherlands, and a member of the board of IOPS, the Dutch Interuniversity Graduate School of Psychometrics and Sociometrics. His research interests include methodological research on the quality of data collected by means of interviewing and the application of the cybernetic approach to the study of social processes. Recent publications are "Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview" (with Maynard, Schaeffer and Houtkoop, 2001) and "Sociocybernetics: Complexity, Autopoiesis, and Observation of Social Systems" (with Geyer, 2001).

Theo van Tilburg is an associate professor in the Department of Sociology and Social Gerontology, Vrije Universiteit, the Netherlands. He is affiliated with the Methods of Data Collection, Living Arrangements and Social Networks of Older Adults, and Longitudinal Aging Study Amsterdam research programs. His research interests include the development and evaluation of survey measuring instruments, the effects of personal network characteristics and social support on well-being, and change in personal networks. His recent publications are "Neighbouring Networks and Environmental Dependency" in Aging & Society *(2000) and "Perceived Instrumental Support Exchanges in Relationships Between Elderly Parents and Their Adult Children" in* Journal of Marriage and the Family *(1999).*